JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 11, 2022

Pp. 1–16
DOI: 10.1089/cmb.2022.0270

Recomb 2022

SCOTv2: Single-Cell Multiomic Alignment with Disproportionate Cell-Type Representation

PINAR DEMETCI,[1,2] REBECCA SANTORELLA,[3] MANAV CHAKRAVARTHY,[2] BJORN SANDSTEDE,[3] and RITAMBHARA SINGH[1,2]

ABSTRACT

Multiomic single-cell data allow us to perform integrated analysis to understand genomic regulation of biological processes. However, most single-cell sequencing assays are performed on separately sampled cell populations, as applying them to the same single-cell is challenging. Existing unsupervised single-cell alignment algorithms have been primarily benchmarked on coassay experiments. Our investigation revealed that these methods do not perform well for noncoassay single-cell experiments when there is disproportionate cell-type representation across measurement domains. Therefore, we extend our previous work—Single Cell alignment using Optimal Transport (SCOT)—by using unbalanced Gromov-Wasserstein optimal transport to handle disproportionate cell-type representation and differing sample sizes across single-cell measurements. Our method, SCOTv2, gives state-of-the-art alignment performance across five non-coassay data sets (simulated and real world). It can also integrate multiple ($M \geq 2$) single-cell measurements while preserving the self-tuning capabilities and computational tractability of its original version.

Keywords: data integration, manifold alignment, multiomics, single-cell sequencing, unbalanced optimal transport.

1. INTRODUCTION

SINGLE-CELL MULTIOMICS MEASUREMENTS provide critical biological insights into cell development and diseases. New experimental protocols have recently been developed for simultaneous measurement of different features in the same single cell (termed ''coassays''). However, no such experimental protocols are available for many feature combinations (e.g., chromatin accessibility and three-dimensional [3D] chromatin conformation) (Clark et al, 2020). Therefore, these features are measured in separately sampled cell populations and analyzed on their own (termed ''non-coassays''). Integrating different measurements from the non-coassays can help explain how different molecular views interact and regulate cellular functions.

[1]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA.
[2]Department of Computer Science, Brown University, Providence, Rhode Island, USA.
[3]Division of Applied Mathematics, Brown University, Providence, Rhode Island, USA.

Unfortunately, these assays lack direct sample–sample and feature–feature correspondences across the measurements. This lack of correspondence makes it hard to use integration methods that require some shared information to perform single-cell alignment (Cao et al, 2020). Therefore, *unsupervised* single-cell multiomics data alignment methods are crucial for integrative single-cell data analysis.

Several unsupervised methods, including our previous work, Single Cell alignment using Optimal Transport (SCOT) (Demetci et al, 2020), have shown competitive performance for integrating different single-cell measurement domains. For example, MMD-MA (Liu et al, 2019; Singh et al, 2020) uses maximum mean discrepancy (MMD) measure to align and embed two data sets in a new space. UnionCom (Cao et al, 2020) performs unsupervised topological alignment through a two-step procedure that first finds a correspondence between the measurement domains, comparing both global and local geometries with a hyperparameter to control the trade-off between them, and then embeds them in a new space. BindSC (Dou et al, 2020) requires the users to bring input data sets to the gene expression feature space by constructing a gene activity score matrix for the epigenomic domains, then finds a correspondence matrix between samples through the bi-order canonical correspondence analysis (bi-CCA), and jointly embeds them into a new space.

Similarly, Seuratv4 (Stuart et al, 2019) also requires gene activity score matrices for epigenomic domains and then identifies correspondence anchors through CCA. Based on these anchors, it imputes one genomic domain based on the other domain and coembeds them into a shared space using UMAP. A recent cross-modal autoencoder-based method (Yang et al, 2021) uses modality-specific autoencoders to map the different modalities to a shared latent space. Our previous method, SCOT (Demetci et al, 2020), compares data from different measurement modalities using Gromov-Wasserstein distances and finds correspondences with optimal transport. Pamona (Cao et al, 2021) extends the SCOT framework with partial Gromov-Wasserstein optimal transport and attempts to identify any cells that do not have existing correspondences in alignment.

A majority of these methods have been mainly evaluated on real-world coassay data sets with underlying 1–1 correspondence between cells across domains. Therefore, our understanding of their performance on non-coassay data sets (with measurements obtained from related but separate cell populations) is limited even though real-world integration tasks involve these data sets. In non-coassay experiments, scientists divide a cell population into smaller aliquots and then apply a different sequencing assay to each. Because the aliquots essentially sample cells from the original population, the resulting data sets can consist of varying proportions of cell types across different measurements, creating cell-type imbalance and lacking 1–1 cell correspondences. In addition, if the original cell population contains rare cell types, it is possible for these cell types to only appear in a subset of these aliquots.

Due to these experimental considerations, we hypothesize that alignment methods that perform well on coassay data sets may not effectively handle the differences in cell-type proportions. Indeed, Pamona (Cao et al, 2021) demonstrated through simulations that current integration methods (Cao et al, 2020; Demetci et al, 2020; Liu et al, 2019; Stuart et al, 2019) tend to perform worse under such settings.

We present SCOTv2, a novel extension of SCOT, that can effectively align both coassay and non-coassay data sets using a single framework. It uses *unbalanced* Gromov-Wasserstein (GW) optimal transport to align data sets with disproportionate cell-type representations while only introducing one additional hyperparameter. This unbalanced framework relaxes the constraint that each point must be mapped with its original mass (a.k.a initially defined marginal probability) during transport. Specifically, an underrepresented cell type in one domain can be transported with more mass to match the proportion of that cell type in the other domains and vice versa. The SCOTv2 framework is summarized in Figure 1. We demonstrate that SCOTv2 aligns data sets with imbalance in cell-type representations better than state-of-the-art baselines. Furthermore, we extend SCOTv2 to integrate single-cell data sets with more than two measurements, making it a multiomics alignment tool.

We perform alignments of six real-world single-cell data sets, with both simulated and natural cell-type imbalance as well as two and more than two domains ($M \geq 2$), demonstrating SCOTv2's applicability across a wide range of scenarios. Finally, similar to the previous version, we present a self-tuning heuristic to select hyperparameters for SCOTv2 without any corresponding information such as cell-type annotations or matching cells or features in truly unsupervised settings.

2. METHODS

Optimal transport finds the most cost-effective way to move data points from one domain to another. One can imagine it as the problem of moving a pile of sand to fill in a hole through the least amount of work.

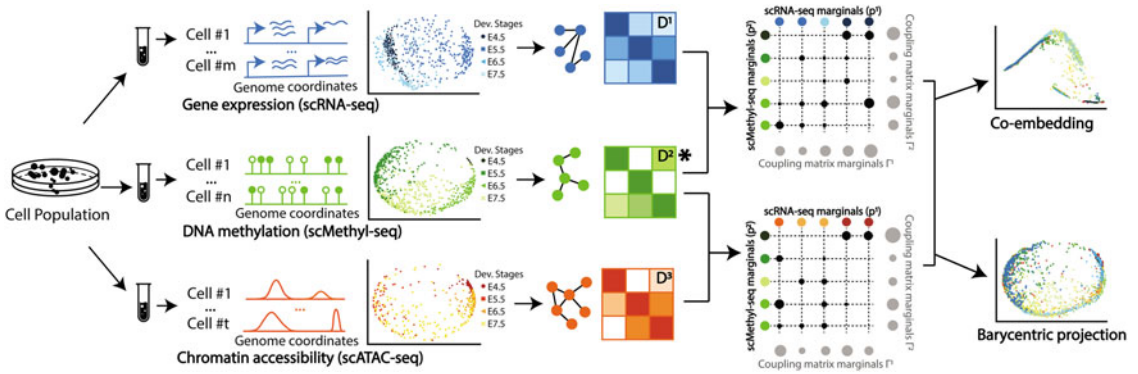

FIG. 1. Overview of SCOTv2 on the scNMT-seq data set (Clark et al, 2018), which contains unbalanced cell-type representation across three domains—RNA expression, chromatin accessibility, and DNA methylation. SCOTv2 selects an anchor domain (denoted with *) and aligns other measurements to it. First, it computes intradomain distance matrices D^m for $m = 1, 2, 3$, which are used to solve for correspondence matrices between the anchor and other domains. The circle sizes in the matrices depict the magnitude of the correspondence probabilities or how much mass to transport. Unbalanced GW relaxes the mass conservation constraint, so the transport map does not need to move each point with its original mass. Finally, it either coembeds the domains into a common space or uses barycentric projections to project them onto the anchor domain. GW, Gromov-Wasserstein.

Our previous framework SCOT (Demetci et al, 2020) uses Gromov-Wasserstein optimal transport, which preserves local geometry when moving data points from one domain to another. The output of SCOT is a matrix of probabilities that represent how likely it is that data points from one modality correspond to data points in the other.

Here, we reintroduce the SCOT formulation to integrate M domains (or single-cell measurements) $X^m = (x_1^m, x_2^m, \ldots x_{n_m}^m) \in \mathbb{R}^{d_m}$ for $m = 1, \ldots M$ with n_m data points (or cells) each. For each data set, we define a marginal distribution p^m, which can be written as an empirical distribution over the data points as follows:

$$p^m = \sum_{i=1}^{n_m} p_i^m \delta_{x_i}. \tag{1}$$

Here, δ_{x_i} is the Dirac measure. For SCOT, we choose these distributions to be uniform over the data.

Gromov-Wasserstein optimal transport performs the transport operation by comparing distances between samples rather than directly comparing the samples themselves (Alvarez-Melis and Jaakkola, 2018). Therefore, for each data set, we compute the intradomain distance matrix D^m. Next, we construct k-NN graphs based on correlations between data points and use Dijkstra's algorithm to compute the shortest path distance on the graph between each pair of nodes. Finally, we connect all the unconnected nodes by the maximum finite distance in the graph and set D^m to be the matrix resulting from normalizing the distances by this maximum.

For two data sets and a given cost function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we compute the fourth-order tensor $\mathbf{L} \in \mathbb{R}^{n_x \times n_x \times n_y \times n_y}$, where $\mathbf{L}_{ijkl} = L(D^1_{ik}, D^2_{jl})$. Intuitively, L quantifies how transporting a pair of points x_i^1, x_k^1 onto another pair across domains, x_j^2, x_l^2, distorts the original intradomain distances and helps to preserve local geometry. Then, the discrete Gromov-Wasserstein problem between p^1 and p^2 is as follows:

$$GW(p^1, p^2) = \min_{\Gamma \in \Pi(p^1, p^2)} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl}, \tag{2}$$

where Γ is a coupling matrix from the set as follws:

$$\Pi(p^1, p^2) = \{\Gamma \in \mathbb{R}_+^{n_1 \times n_2} : \Gamma 1_{n_2} = p_1, \Gamma^T 1_{n_1} = p_2\}. \tag{3}$$

One of the advantages of using optimal transport is the probabilistic interpretation of the resulting coupling matrix Γ, where the entries of the normalized row $\frac{1}{p_i}\Gamma_i$ are the probabilities that the fixed data point x_i corresponds to each y_j. Each entry Γ_{ij} describes how much of the mass of x_i should be mapped to y_j.

To make this problem more computationally tractable, we solve the entropically regularized version as follows:

$$GW_\varepsilon(p^1, p^2) = \min_{\Gamma \in \Pi(p^1, p^2)} \langle \mathbf{L}(D^1, D^2) \otimes \Gamma, \Gamma \rangle - \varepsilon H(\Gamma). \tag{4}$$

where $\varepsilon > 0$ and $H(\Gamma)$ is the Shannon entropy defined as $H(\Gamma) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \Gamma_{ij} \log \Gamma_{ij}$. Larger values of ε make the problem more convex but also lead to a denser coupling matrix, meaning there are more correspondences between samples. In SCOT, we use the cost function $L = L_2$.

2.1. Unbalanced optimal transport of SCOTv2

Our proposed solution to align data sets with different numbers of samples or proportions of cell types is to use unbalanced optimal transport, which adds divergence terms to allow for mass variations in the marginals (Liero et al, 2018; Sjourn et al, 2021). We follow Sjourn et al (2021) and use the Kullback–Leibler divergence as follows:

$$\mathrm{KL}(p||q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right), \tag{5}$$

to measure the difference between the marginals of the coupling Γ and the input marginals p^1 and p^2. Thus, we solve the unbalanced GW problem:

$$GW_{\varepsilon,\rho}(p^1, p^2) = \min_{\Gamma \geq 0} \langle \mathbf{L}(D^1, D^2) \otimes \Gamma, \Gamma \rangle - \varepsilon H(\Gamma) + \rho \mathrm{KL}(\Gamma 1_{n_2} || p^1) + \rho \mathrm{KL}(\Gamma^T 1_{n_1} || p^2), \tag{6}$$

where $\rho > 0$ is a hyperparameter that controls the marginal relaxation. When ρ is large, the marginals of Γ should be close to p^1 and p^2, and when ρ is small, the marginals of Γ may differ more, allowing each point to transport with more or less mass than it originally had. See Algorithm 1 for details.

2.2. Extending SCOTv2 for multidomain alignment

To align more than two data sets ($M > 2$), we use one domain as an anchor to align the other domains. The anchor should be the domain with the clearest biological structures, for example, a data set with the best-defined cell-type clusters. We propose selecting the anchor through the kNN graph used to compute D^m. For every node x_i^m in the graph, we calculate the average of the k neighboring node values $\mathcal{N}_k(x_i^m)$. Next, we measure the difference between this average and the true value of the node. This difference reflects how well the averaged neighborhood represents the given node. We then average these differences across the graph and select the domain with the lowest averaged difference as the anchor. Intuitively, we select the anchor whose kNN graph best reflects its data set. Suppose X^1 is the anchor data set. Then, for $m = 2, 3, \ldots, N$, we compute the coupling matrix Γ^m according to Equation (4).

Algorithm 1: Pseudocode for Unbalanced GW Optimal Transport (UGWOT)

Input: Marginal probabilities p^1 and p^2, intradomain distance matrices D^1 and D^2, relaxation coefficient ρ, regularization coefficient ϵ
Initialize the coupling matrix: $\Gamma = \pi = p^1 \otimes p^2$
while Γ *not converged* **do**
  $\Gamma \leftarrow \pi$
  $\Gamma_{(mass)} \leftarrow \sum_{i,j} \Gamma_{i,j}$ $\tilde{\epsilon} \leftarrow \Gamma_{(mass)}\epsilon$, $\tilde{\rho} \leftarrow \Gamma_{(mass)}\rho$
  //Compute cost C:
  $\Gamma^1 \leftarrow \Gamma 1_{n_2}$, $\Gamma^2 \leftarrow \Gamma^T 1_{n_1}$
  $A \leftarrow (D^1)^{\circ 2}\Gamma^1$, $B \leftarrow (D^2)^{\circ 2}\Gamma^2$
  $D \leftarrow D^1 \Gamma D^2$
  $E \leftarrow \epsilon \sum_{ij} \log\left(\frac{\Gamma_{i,j}}{p_i^1 p_j^2}\right)\Gamma_{i,j} + \rho\left(\sum_i \log\left(\frac{\Gamma_i^1}{p_i^1}\right)\Gamma_i^1 + \sum_j \log\left(\frac{\Gamma_j^2}{p_j^2}\right)\Gamma_j^2\right)$
  $C \leftarrow A + B - 2D + E$
  //Perform Sinkhorn iterations
  while (u, v) *not converged* **do**
    $u \leftarrow -\frac{\tilde{\epsilon}\tilde{\rho}}{\tilde{\epsilon}+\tilde{\rho}} \log\left[\sum_{i,j} \exp\left(v_j - C_{ij}\right)/\tilde{\epsilon} + \log p^2\right]$
    $v \leftarrow -\frac{\tilde{\epsilon}\tilde{\rho}}{\tilde{\epsilon}+\tilde{\rho}} \log\left[\sum_{i,j} \exp\left(u_i - C_{ij}\right)/\tilde{\epsilon} + \log p^1\right]$
  end
  //Update: $\pi_{ij} \leftarrow \exp\left[u_i + v_j - C_{ij}\right] p_i^1 p_j^2$
  //Rescale: $\pi \leftarrow \sqrt{\Gamma_{(mass)}/\pi_{(mass)}}\pi$
end
Return: Γ

To have all of the data sets aligned in the same domain, we can either use barycentric projection to project each X^m for $m=2, 3, \ldots, M$ onto X^1 or find a shared embedding space as described in Section 2.3. In the first iteration of SCOT, we used a barycentric projection to align and project one data set onto the other. Due to the marginal relaxation, we now search for a non-negative $n_1 \times n_m$ dimensional matrix Γ instead of $\Gamma \in \Pi(p^1, p^m)$. Because of this change, the adjusted barycentric projection is as follows:

$$x_i^m \mapsto \frac{\sum_{j=1}^{n_1} \Gamma_{ij}^m x_j^1}{\sum_{j=1}^{n_1} \Gamma_{ij}^m}. \tag{7}$$

2.3. Embedding with the coupling matrix

Other methods such as MMD-MA and UnionCom align data sets by embedding them into a common latent space of dimension $p \leq \min_{m=1,\ldots,M} d_m$. Here d_m represents the original dimension size of measurement (or domain) m. Embedding the data sets in a new space often leads to a better alignment as it introduces the additional benefits of dimensionality reduction, allowing more meaningful structures in the data sets, such as cell-type clusters, to be more prevalent. Due to these benefits, we also enable the embedding option through a modification of the t-Distributed Stochastic Neighbor Embedding (t-SNE) method (Van der Maaten and Hinton, 2008) as proposed by UnionCom (Cao et al, 2020).

For each domain m, we compute P^m, an $n_m \times n_m$ cell-to-cell transition matrix; each entry $P_{j|i}^m$ is the conditional probability that a data point x_i^m would pick x_j^m as its neighbor when chosen according to a Gaussian distribution centered at x_i^m:

$$P_{j|i}^m = \frac{\exp\left(-||x_i^m - x_j^m||^2/2\sigma_i^2\right)}{\sum_{k\neq i} \exp\left(-||x_i^m - x_k^m||^2/2\sigma_i^2\right)}. \tag{8}$$

Here, the bandwidth σ_i is chosen according to the density of the data points through a binary search for the value of σ_i that achieves the user-supplied perplexity value. P^m is computed by averaging $P_{i|j}^m$ and $P_{j|i}^m$ to give more weight to outlier points:

$$P_{ij}^m = \frac{P_{i|j}^m + P_{j|i}^m}{2n_m} \tag{9}$$

Similarly, for the lower dimensional embeddings, we compute a cell-to-cell probability matrix $Q^{m'}$ through a Student-t distribution with one degree of freedom:

$$Q_{ij}^{m'} = \frac{\left(1 + ||x_i^{m'} - x_j^{m'}||\right)^{-1}}{\sum_{k\neq l} 1 + \left(||x_k^{m'} - x_l^{m'}||\right)^{-1}}. \tag{10}$$

Then, to jointly embed all domains through the anchor domain X^1, the optimization problem is as follows:

$$\min_{X^{1'},\ldots,X^{M'}} \sum_{m=1}^{M} \mathrm{KL}(P^m||Q^{m'}) + \beta \sum_{m=2}^{M} ||X^{1'} - X^{m'}(\Gamma^m)^T||_F^2, \tag{11}$$

where $X^{m'}$ is the lower dimensional embedding of X^m, and Γ^m is the coupling matrix from solving Equation (6) for $m=2, \ldots, M$. These two terms seek to find an embedding that both preserves the local geometry in the original domain and aligns the domains according to the correspondence found by GW. The intuition behind the term $\mathrm{KL}(P^m||Q^{m'})$ is very similar to that of GW; if two points have a high transition probability in the original space, then they should also have a high transition probability in the latent space. The term $||X^{1'} - X^{m'}(\Gamma^m)^T||_F^2$ measures how well aligned the new embeddings $X^{1'}$ and $X^{m'}$ are according to the prescribed coupling matrix Γ^m. Finally, $\beta > 0$ controls the trade-off between preserving the original geometry with the KL term and enforcing the alignment found with GW. We solve this optimization problem using a gradient descent from UnionCom with a default latent space dimension size $p=3$ (Cao et al, 2020). The overall SCOTv2 method is presented as Algorithm 2.

Algorithm 2: Pseudocode for SCOTv2 Algorithm

Input: Data sets $X^1, \ldots, X^M$, number of neighbors in nearest neighbor graphs k, entropic regularization coefficient ε, mass conservation relaxation coefficient ρ.

for $m = 1, \ldots, M$ **do**

  //Initialize marginal probabilities: $p^m \leftarrow \text{Uniform}(X^m)$;

  //Construct G^m, a $k-$NN graph based on pairwise correlations

  //Compute intradomain distance matrix D^m on G^m with Dijkstra's algorithm.

  //Compute a ''neighborhood correlation'' score, c^m:

  $c^m = \frac{1}{n_m} \sum_{i=1}^{n_m} \frac{1}{k} \sum_{x_j^m \in \mathcal{N}_k(x_i^m)} \text{corr}(x_j^m, x_i^m)$

end

//Select an anchor domain X^{m*}: $m^* = \arg\max_{m=1,\ldots M} c^m$

for $m = 1, \ldots, M$ $(m \neq m^*)$ **do**

  //Compute pairwise coupling matrices between the anchor domain X^{m*} and all other domains:

  $\Gamma^m \leftarrow GW_{\varepsilon,\rho}(p^m, p^{m*})$

  if *Barycentric projection* **then**

    $x_i^{m'} \leftarrow \frac{\sum_{j=1}^{n_1} \Gamma_{ij}^m x_j^{m*}}{\sum_{j=1}^{n_1} \Gamma_{ij}^m}$

  end

  else

    //Find shared embedding

    $X^{1'} \ldots X^{M'} \leftarrow \min_{X^{m'}, \ldots, X^{M'}} \sum_{m=1}^{M} \text{KL}(P^m || Q^{m'})$

    $+ \beta \sum_{m \neq m*} ||X^{m^{*'}} - X^{m'}(\Gamma^m)^T||_F^2$

  end

end

Return: Aligned data sets, $X^{1'} \ldots X^{M'}$.

Algorithm 3: Unsupervised hyperparameter search procedure

Input: Data sets $X^1, \ldots, X^M$.

//Find k for each domain

for $m = 1, \ldots, M$ **do**

  $k^m = \underset{k \in \{10, 20, \ldots, 150\}}{\text{argmax}} \frac{1}{n_m} \sum_{i=1}^{n_m} \frac{1}{k} \sum_{x_j^m \in \mathcal{N}_k(x_i^m)} \text{corr}(x_j^m, x_i^m)$

  //Use k^m to compute D^m

end

//Use the GW distance to pick ρ and ϵ

for $m = 2, \ldots, M$ **do**

  $\epsilon^m, \rho^m = \arg\min_{\epsilon,\rho} GW_{\epsilon,\rho}(\mathbb{1}_{n_1}, \mathbb{1}_{n_m})$

end

Return: k^m, ϵ^m, ρ^m.

2.4. Heuristic for self-tuning hyperparameters

SCOTv2 has three hyperparameters: (1) k for the number of neighbors to consider in nearest neighbor graphs, (2) the weight of the entropic regularization term, ϵ, and (3) the coefficient of the mass relaxation constraint, ρ. The barycentric projection of one domain onto another does not require any hyperparameters. However, jointly embedding the domains in a latent space requires selecting the dimension p.

Ideally, orthogonal correspondence information such as 1–1 correspondences and cell-type labels can guide hyperparameter tuning as validation. However, such information is hard to obtain in most cases. First, no validation data on cell-to-cell correspondences exist for non-coassay data sets. Second, it is challenging to infer cell types for certain sequencing domains such as 3D chromatin conformation. Lastly, the cell-type annotations may not always agree across single-cell domains.

We provide a heuristic to self-tune hyperparameters in a completely unsupervised setting. We first choose a k for the neighborhood graphs that yields a high average correlation value between the neighborhood predicted values and measured genomic values of the graph nodes. This step is the same as the one used to select the anchor domain for multiomics alignment in Section 2.2. Next, we choose ϵ and ρ values that minimize the Gromov-Wasserstein distance between the aligned data sets. Algorithm 3 gives the details of this procedure.

3. EXPERIMENTAL SETUP

3.1. Data sets

We evaluate SCOTv2 on single-cell data sets with disproportionate cell types using two schemes. (1) We subsample different cell types in coassay data sets to simulate cell-type representation disparities between sequencing modalities. (2) We select real-world single-cell data sets with separately assayed measurement modalities, which lack 1–1 cell correspondences and have different cell-type proportions across modalities due to the fact that they were profiled on separate cell populations. In addition, we present results on the original coassay data sets with 1–1 cell correspondence to demonstrate the flexibility of SCOTv2 across balanced and unbalanced single-cell data sets.

3.1.1. Coassay single-cell data sets with 1–1 cell correspondence. We use three coassay data sets to validate our model, sequenced by SNARE-seq, scGEM, and scNMT-seq technologies. SNARE-seq is a two-modality sequencing technology that simultaneously captures the chromatin accessibility and transcriptional profiles of cells (Chen et al, 2019). This data set contains a total of 1047 cells from four cell lines: BJ (human fibroblast cells), H1 (human embryonic cells), K562 (human erythroleukemia cells), and GM12878 (human lymphoblastoid cells; Gene Expression Omnibus access code: GSE126074). We follow the same data preprocessing steps outlined by Chen et al (2019) and work with the top 10 principal components of the gene expression domain and the 19 topics selected by cisTopic (González-Blas et al, 2018) in the chromatin accessibility domain.

The scGEM technology is a three-modality sequencing technology that profiles the genetic sequence, gene expression, and DNA methylation states in the same cell (Cheow et al, 2016). The data set we use is derived from human somatic cell samples undergoing conversion to induced pluripotent stem cells (Sequence Read Archive accession code SRP077853) (Cheow et al, 2016). We access the preprocessed data provided by Welch et al (2017), which only contain the gene expression and DNA methylation modalities.* The data set sequenced by the scNMT-seq method (Argelaguet et al, 2019) contains three modalities of genomic data: gene expression, DNA methylation, and chromatin accessibility, from mouse gastrulation samples, going through the Carnegie stages of vertebrate development (Gene Expression Omnibus access code: GSE109262).

We access the preprocessed data through the Bioconductor package named ''Mouse Gastrulation Data,''† which was released by the authors. They also provide the scripts they use to preprocess the raw data.‡ While the SNARE-seq and scGEM data sets contain the same number of cells across measurements, scNMT-seq modalities contain different cell-type proportions after preprocessing due to varying noise levels in measurements. Table 1 lists the number of cells belonging to different cell types in each domain for the scNMT-seq data set.

3.1.2. Single-cell data sets with simulated cell-type imbalance. To test alignment performance sensitivity to different levels and types of cell-type proportion disparities across modalities, we generate simulation data sets by subsampling SNARE-seq and scGEM cosequencing data sets in three ways. (1) We remove a cell type from one modality. (2) We reduce the proportion of a cell type in one modality by subsampling it at 50% and another cell type in the other modality by subsampling it at 75%. We simulate

*Preprocessed data for the scGEM data set accessed here: https://github.com/jw156605/MATCHER

†''Mouse Gastrulation Data'' from scNMT-seq accessed through Bioconductor by following these steps: https://bioconductor.org/packages/release/data/experiment/vignettes/MouseGastrulationData/inst/doc/MouseGastrulationData.html

‡Preprocessing scripts for the scNMT-seq data accessed here: https://github.com/PMBio/scNMT-seq

Table 1. Number of Cells in (and Percentages of) Each Cell Type Across Different Modalities in the scNMT-seq Coassayed Data Set After Quality Control Procedures and the Non-Coassay Data Sets

	Modality 1 (gene expression)	*Modality 2 (chromatin accessibility or chromatin imaging)*	*Modality 3 (DNA methylation or 3D chromosomal conformation)*
scNMT data set	$n = 579$	$n = 647$	$n = 725$
	E4.5: 76 (12.73%)	E4.5: 63 (9.73%)	E4.5: 65 (8.96%)
	E5.5: 104 (17.42%)	E5.5: 89 (13.76%)	E5.5: 91 (12.55%)
	Day 6.5: 146 (24.46%)	E6.5: 220 (34.00%)	E6.5: 278 (38.34%)
	E7.5: 271 (45.39%)	E7.5: 175 (42.50%)	E7.5: 291 (40.14%)
sciOmics data set	$n = 1058$	$n = 1296$	$n = 2154$
	Day 0: 489 (46.22%)	Day 0: 164 (12.65%)	Day 0: 987 (45.82%)
	Day 3: 127 (12.00%)	Day 3: 702 (54.17%)	Day 3: 435 (20.19%)
	Day 7: 78 (7.37%)	Day 7: 77 (5.94%)	Day 7: 243 (11.28%)
	Day 11: 145 (13.71%)	Day 11: 175 (13.50%)	Day 11: 164 (7.61%)
	NPC: 219 (20.70%)	NPC: 178 (13.73%)	NPC: 325 (15.09%)
MEC data set	$n = 26{,}273$	$n = 21{,}262$	
	Basal: 11,138 (42.39%)	Basal: 13,353 (62.80%)	
	L-Sec (Prog): 7683 (29.24%)	L-Sec (Prog): 3343 (15.72%)	
	L-HR: 3439 (13.09%)	L-HR: 2624 (12.34%)	N/A
	L-Sec (Mat): 2869 (10.92%)	L-Sec (Mat): 1165 (5.48%)	
	L-Sec (Prolif): 758 (2.89%)	L-Sec (Prolif): 7 (0.033%)	
	Stroma: 386 (1.47%)	Stroma: 770 (3.62%)	
RNA-imaging data set	$n = 1166$	$n = 3482$	
	Quiescent T cells: 545 (46.74%)	Quiescent T cells: 343 (9.85%)	N/A
	Poised T cells: 621 (53.26%)	Poised T cells: 3139 (90.15%)	

3D, three dimensional; MEC, mammary epithelial cell; NPC, neural progenitor cell.

this setting to test how the alignment methods will behave when multiple cell types have disproportionate representation at different levels (e.g., half or quarter percentage of cell types missing) across modalities. (3) We randomly pick a modality and downsample it at one of the following five downsampling rates: 10%, 20%, 30%, 40%, *and* 50% to test the robustness of alignment algorithms to varying downsampling rates.

For these cases, we uniformly pick at random which cell type to subsample or remove. Specifically, for scGEM in simulation case (1), we remove ''d16T+'' cells in the DNA methylation domain while retaining the original gene expression domain, and remove the ''d24T+'' cells in the gene expression domain while retaining the original DNA methylation domain. For the SNARE-seq data set, we remove ''GM'' cells in the gene expression domain and ''K562'' in the chromatin accessibility domain. In simulation case (2), we subsample the ''d8'' cluster of the scGEM data set at 75% in the gene expression modality and the ''d16T+'' cluster at 50% in the DNA methylation modality. For SNARE-seq, we subsample the ''H1'' cluster at 75% and the ''K562'' cluster at 50% in the gene expression and chromatin accessibility domains, respectively.

3.1.3. Single-cell data sets without 1–1 correspondences. We also align non-coassay data sets containing separately sequenced single-cell measurements. Bonora et al (2021) generated the first data set we use, which we call ''sciOmics.'' This data set consists of sciRNA-seq, sciATAC-seq, and sciHiC measurements, capturing gene expression, chromatin accessibility, and 3D chromosomal conformation profiles of mouse embryonic stem cells undergoing differentiation. The measurements were taken at the following five stages: days 0, 3, 7, 11, and as fully differentiated neural progenitor cells (NPCs). The second non-coassay data set, ''MEC,'' contains gene expression and chromatin accessibility measurements taken using the 10X Chromium scRNA-seq and scATAC-seq technologies on mouse mammary epithelial cells (MECs). Since each modality consists of separately sampled cell populations, these contain disparate cell-type proportions across modalities.

Lastly, to demonstrate the general applicability of SCOTv2 to measurement modalities beyond sequencing data sets, we align unpaired single-cell chromatin images and gene expression profiles of T cells

from human peripheral blood (''RNA-imaging'' data set) (Yang et al, 2021). We accessed this data set through the GitHub repository https://github.com/uhlerlab/cross-modal-autoencoders For the imaging data set, we use the tensor representation provided by the data loader in the repository.

Table 1 lists the number of cells belonging to different cell types in each domain for the sciOmics, MEC, and RNA-imaging data sets.

3.2. Evaluation metrics and baseline methods

Although most of the data sets lack 1–1 cell correspondences, we can evaluate alignment using cell-type labels through label transfer accuracy (LTA) as in Cao et al (2021), Cao et al (2020), and Demetci et al (2020). This metric assesses the clustering of cell types after alignment by training a *k*NN classifier on a training set (50% of the aligned data) and then evaluates its predictive accuracy on a test data set (the other 50% of the aligned data). Higher values correspond to better alignments, indicating that cells that belong to the same cell type are aligned close together after integration.

We benchmark our method against the current unsupervised single-cell multiomic alignment methods outlined in Section 1, namely, MMD-MA (Singh et al, 2020), UnionCom (Cao et al, 2020), bindSC (Dou et al, 2020), Seuratv4 (Stuart et al, 2019), Pamona (Cao et al, 2021), cross-modal autoencoders (Yang et al, 2021), and the previous version of SCOT (Demetci et al, 2020), which performs alignment without the KL term. For each of these benchmarks, we define a hyperparameter grid of similar granularity and perform extensive tuning, as detailed in Section 3.3.

Both Pamona and cross-modal autoencoders allow the users to provide weak supervision to their algorithms using the cell-type labels. Aligning data sets using the cell-type annotations lead to better alignment outcomes. However, we consider the unsupervised integration scenario in this article since annotating cell types for measurement modalities other than gene expression can be challenging and lead to contradicting labels across domains. As a result, we do not use this optional feature in Pamona and remove the cell-type loss term from the overall loss in the scripts provided by cross-modal autoencoders. In addition, of all the selected baselines, only Pamona and UnionCom provide a way to align more than two domains at once, so we only use them as baselines for experiments with multiple domains ($M > 2$).

3.3. Hyperparameter tuning

For each data set and alignment method, we report results with the best performing hyperparameter combination in Section 4.1. When defining hyperparameter search grids, if methods share similar hyperparameters in their formulation, we keep the range defined for these consistent across all algorithms. Examples for such hyperparameters are dimensionality of the latent space, p, for the algorithms that commonly embed data sets; entropic regularization constant, ϵ, for methods that use optimal transport; and number of neighbors, k, for methods that model single-cell data sets with nearest neighbor graphs. Otherwise, we refer to the publication and the code repository for each method to choose a hyperparameter range.

For Pamona, we tune the following four hyperparameters: $k \in \{20, 30, \ldots, 150\}$, the number of neighbors in the cell neighborhood graphs, $\epsilon \in \{5e-4, 3e-4, 1e-4, 7e-3, 5e-3, \ldots, 1e-2\}$, the entropic regularization coefficient for the optimal transport formulation, $\lambda \in \{0.1, 0.5, 1, 5, 10\}$, the coefficient for the trade-off between aligning corresponding cells and preserving local geometries, and lastly, $p \in \{3, 4, 5, 10, 30, 32\}$, the output dimension for embedding. We choose the ranges for ϵ and k to be consistent with the corresponding hyperparameters in the SCOT and SCOTv2 algorithms and the ranges for the embedding dimensions to be consistent with the recommended values in the MMD-MA and UnionCom embeddings.

For UnionCom, we tune the trade-off parameter $\beta \in \{0.1, 1, 5, 10, 15, 20\}$ and the regularization coefficient $\rho \in \{0, 0.1, 1, 5, 10, 15, 20\}$ based on the ranges reported by Cao et al (2020) in the publication. We additionally tune the maximum neighborhood size permitted in the neighborhood graphs, $k_{max} \in \{40, 100, 150\}$, as well as the embedding dimensionality $p \in \{3, 4, 5, 10, 30, 32\}$. The sweep range for hyperparameter k_{max} is smaller than the other hyperparameters because UnionCom automatically starts from $k = 2$ and goes up to k_{max} to find the lowest k that returns a connected graph to use in the algorithm. Therefore, more refined search is not needed.

For MMD-MA, we choose the weights λ_1 and $\lambda_2 \in \{1e-2, 5e-3, 1e-3, 5e-4, \ldots, 1e-9\}$. This range includes the hyperparameter range suggested by Singh et al ($\lambda_1, \lambda_2 \in \{1e-3, 1e-4, 1e-5, 1e-6, 1e-7\}$)

but extends it further to increase the granularity for the sake of more fair comparison against methods that require a higher number of hyperparameters to test, such as Pamona and UnionCom. Similarly to other methods, we also select the embedding dimensionality from $p \in \{3, 4, 5, 10, 30, 32\}$.

For bindSC, we choose the couple coefficient that assigns weight to the initial gene activity matrix $\alpha \in \{0, 0.1, 0.2, \ldots 0.9\}$ and the couple coefficient that assigns weight factor to multiobjective function $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$. In addition, we choose the number of canonical vectors for the embedding space $K \in \{3, 4, 5, 10, 30, 32\}$. For Seurat v4, we tune the number of neighbors to consider when finding anchors, $k \in \{5, 10, 15, 20\}$, dimensions of the final coembedding space, $p \in \{3, 4, 5, 10, 30, 32\}$, and the choice of the reference and anchor domains when finding anchors.

Lastly, for cross-modal autoencoders, we tune the hyperparameters that control how well the data distribution in each domain will be preserved in the latent space, $\lambda_1 \in \{1e-8, 1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1\}$ for the first domain, and $\lambda_2 \in \{1e-8, 1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1\}$ for the second domain. We also consider the following values for the size of the latent space where the aligned data are embedded, $p \in \{3, 4, 10, 30, 32, 128, 256, 512\}$.

4. RESULTS

4.1. SCOTv2 gives high-quality alignments consistently across all single data sets

We first present the alignment results for real-world coassay data sets with simulated cell-type imbalance. We present the results obtained by the best performing hyperparameter combinations for all methods compared in this study. Figure 2A visualizes the barycentric projection alignments performed by SCOTv2, with the first two, principal components (PC) plotted for the SNARE-seq and scGEM data sets, respectively. We use barycentric projection for visualization purposes for the ease of comparison with the original domains, plotted in Supplementary Figure S1.

Here, we integrate data sets under three different settings described in the previous section: (1) balanced data sets (or ''full data sets'' with no subsampling), (2) missing cell types in the epigenomic domains, and (3) subsampled cells in both domains (one cell type at 50% in the epigenomic domains and another cell type at 75% in the gene expression domains). We include alignment results on the full data sets with 1–1 sample correspondences to ensure that SCOTv2 performs well for balanced cases as well.

Qualitatively, we see that SCOTv2 preserves the cell-type annotations after alignment for all three settings. In Figure 2B, we report the quantitative performance of SCOTv2 and all the other state-of-the-art baselines using the LTA scores. MMD-MA, UnionCom, Seurat, and bindSC fail to reliably align data sets with disproportionate cell-type representation across modalities. While Pamona tends to yield high-quality alignments for cases with cell-type disproportion, it fails to perform well on the SNARE-seq balanced data set as well as its subsampling simulation. We additionally apply Pamona to randomly downsampled coassays (Fig. 3). We show that while Pamona's partial optimal transport framework handles cell-type disproportion better than the balanced optimal transport formulation (demonstrated by SCOT), SCOTv2 still shows an advantage in all SNARE-seq simulations ($\sim$20% increase in LTA), as well as the smaller downsampling schemes ($\sim$10%).

Among all the methods tested, SCOTv2 consistently gives more high-quality alignments across different scenarios of cell-type representation. It also demonstrates a $\sim$22% average increase in LTA over the previous version of the algorithm (SCOT) when comparing the barycentric projection results and $\sim$27% for the embedding results. Figure 3 presents similar results (SCOTv2 attains an LTA of 0.786 followed by Pamona at 0.62 on SNAREseq and 0.542 followed by Pamona at 0.538 on scGEM) for missing cell types in the other (gene expression) domain, suggesting that our choice of domain with missing cell type does not affect the performance comparison results. UnionCom, Pamona, and SCOTv2 allow us to perform both barycentric projections and embed the single-cell domains in a lower dimensional space.

Overall, we observe that embedding yields higher LTA values than barycentric projection. Since the barycentric projection projects one domain onto another, the separation of the domain being projected onto (or anchor domain) limits the clustering separation after alignment. In contrast, the embedding utilizes t-SNE to enhance cell-type separation, allowing for better-separated clusters after alignment.

Next, we report the alignment performance of SCOTv2 on single-cell data sets with inherent disparities in cell-type representation, mostly due to sampling during experiments. We include scNMT, a coassay with varying levels of cells across domains due to quality control procedures, along with sciOmics, MEC, and

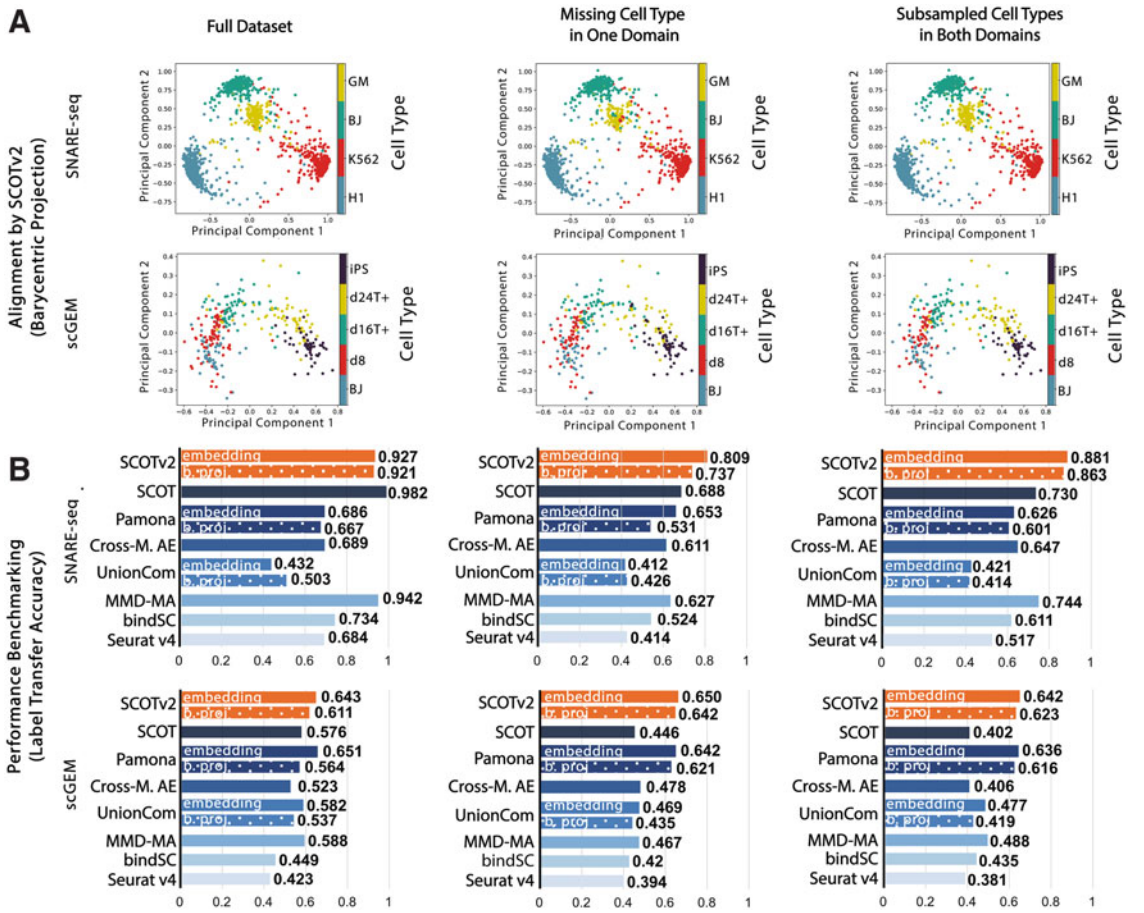

FIG. 2. Alignment results for simulations and balanced coassay data sets. **(A)** Visualizes the barycentric projection alignment on SNARE-seq and scGEM for the full coassay data sets, simulations with a missing cell type in the epigenomic domain, and subsampled cell types in both domains. **(B)** Compares the alignment performance of SCOTv2 with the benchmarks through LTA. For SCOTvs, Pamona, and UnionCom, we report results on both embedding into a shared space (solid bars) and the barycentric projection (dotted bars). LTA, label transfer accuracy.

RNA-imaging data sets for this experiment. Note that scNMT and sciOmics have three different modalities, and hence, we can only report the baselines for methods that provide a way to align data sets with $M > 2$ in their released code. Figure 4A presents the qualitative alignment results for SCOTv2 (for visualization purposes, the aligned data sets are reduced to a 2-dimensional space through PCA). SCOTv2 performs well on all real-world data sets with disproportionate cell-type representation across modalities, including the ones with three modalities.

The LTA scores in Figure 4B demonstrate that SCOTv2 consistently yields the best alignments on the four real-world data sets. These results highlight its ability to reliably integrate separately sampled data sets with disproportionate cell-type representation and multiple ($M > 2$) modalities simultaneously.

4.2. Hyperparameter self-tuning aligns well without depending on orthogonal correspondence information

The benchmarking results above present the alignment performance of each algorithm at its best hyperparameter setting; however, users may not have 1–1 correspondences to validate alignments, for the purpose of hyperparameter selection, in real-world applications. While users may have access to cell-type labels, inferring that cell types are highly difficult in specific modalities of single-cell sequencing, such as 3D chromatin conformation. In addition, different sequencing modalities might disagree on cell-type clustering (as is often the case with the scRNA-seq and scATAC-seq data sets). In these situations, users might not have sufficient validation data for tuning hyperparameters.

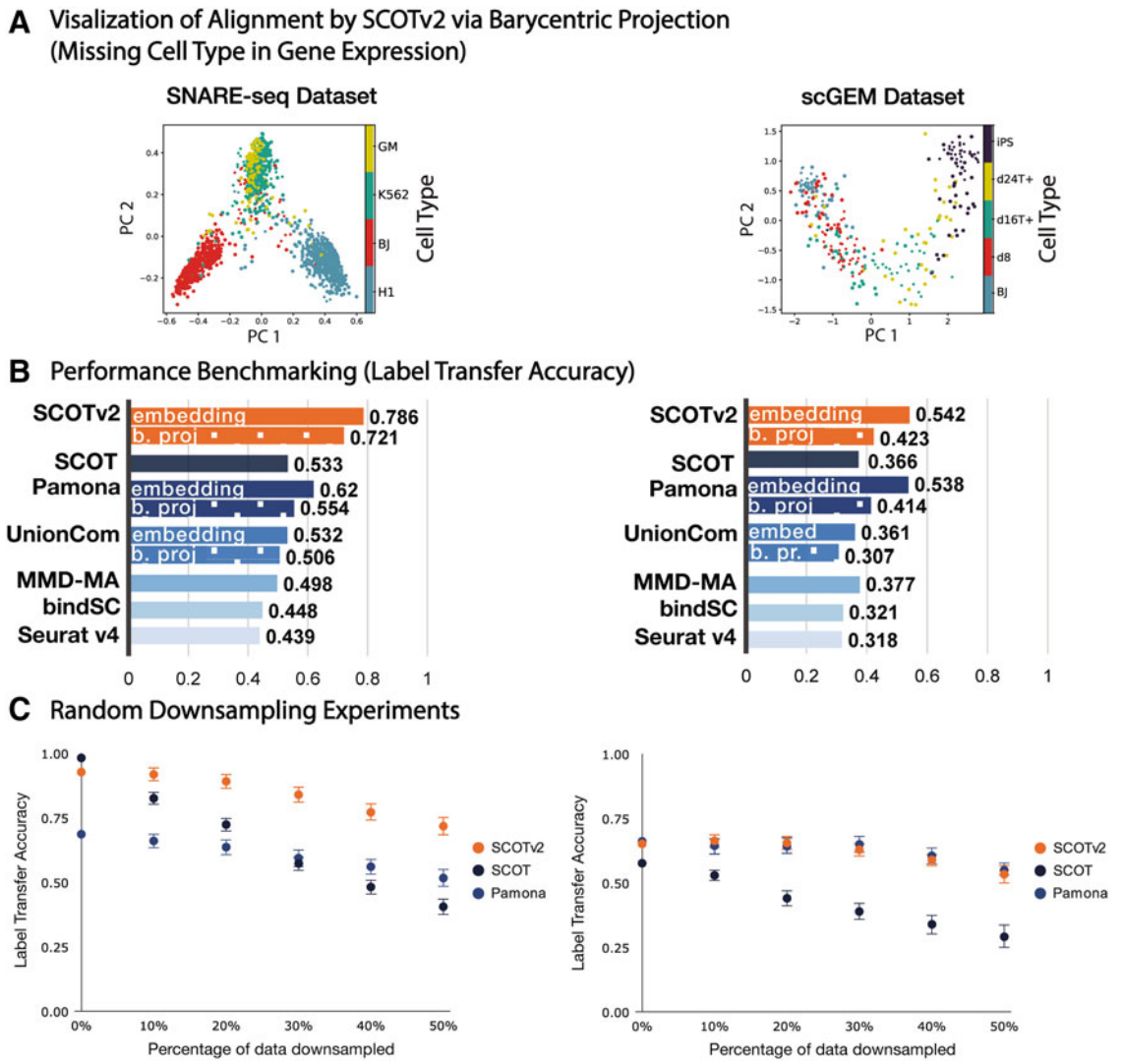

FIG. 3. Alignment performance under varying subsampling ratios. **(A)** Presents results on SNARE-seq data set and **(B)** presents results on scGEM data set. **(C)** For both, we randomly subsample cells at different subsampling ratios and perform alignment with SCOT, SCOTv2, and Pamona. Each case is repeated five times. SCOT, Single Cell alignment using Optimal Transport.

We design a heuristic process (described in Section 2.4), as done previously for SCOT, that allows SCOTv2 to select hyperparameters in a completely unsupervised manner. Other alignment methods do not provide an unsupervised hyperparameter tuning procedure. Therefore, without validation data, a user would have to use the default parameters. In Table 2, we compare the alignment performance for our heuristic against the default parameters of other methods. While our heuristic does not always yield the optimal hyperparameter combination, it does give more favorable results over the default settings of the other methods. Thus, we recommend using it in cases that lack orthogonal information for hyperparameter tuning.

4.3. SCOTv2 scales well with increasing number of samples

We compare the runtime of SCOTv2 with the top performing methods: Pamona, MMD-MA, UnionCom, and the previous version of SCOT by subsampling various numbers of cells from the MEC data set. MMD-MA, UnionCom, and SCOTv2 have GPU versions, while Pamona and SCOT only have CPU versions. We run MMD-MA and UnionCom on a single NVIDIA GTX 1080ti GPU with VRAM of 11 GB and Pamona

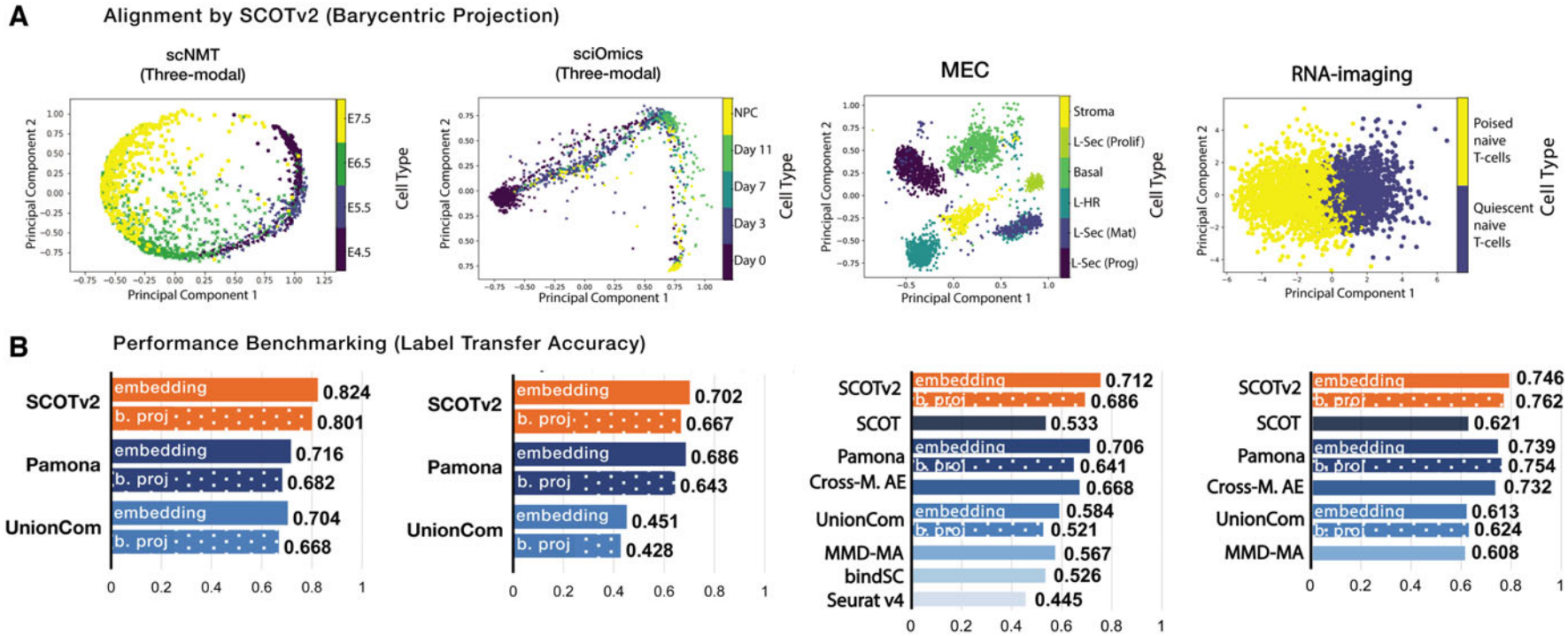

FIG. 4. Alignment results for multimodal ($M > 2$) and separately sequenced data sets. **(A)** Visualizes the alignment of scNMT-seq, sciOmics, and MEC. All data sets have unequal sample sizes and cell-type proportions across domains. **(B)** Benchmarks alignment performance through LTA. As in Figure 2, we report results both by embedding (solid bars) and barycentric projection (dotted bars) for the methods that allow for both. For scNMT-seq and sciOmics, which are three-modal data sets, we only demonstrate results for SCOTv2, Pamona, and UnionCom, which can handle more than two modalities. MEC, mammary epithelial cell.

Table 2. Alignment Performance Benchmarking in the Fully Unsupervised Setting

	SNARE (full data set)	*SNARE (missing cell type)*	*SNARE (subsam.data set)*	*scGEM (full data set)*	*scGEM (missing cell type)*	*scGEM (subsam.data set)*	*scNMT*	*sciOmics*	*MEC*
SCOTv2	0.826	**0.653**	**0.751**	**0.509**	**0.521**	**0.415**	**0.727**	**0.537**	**0.584**
SCOT	**0.852**	0.572	0.588	0.423	0.323	0.314	N/A	N/A	0.466
Pamona	0.554	0.423	0.419	0.385	0.414	0.308	0.588	0.329	0.417
MMD-MA	0.523	0.407	0.431	0.360	0.296	0.287	N/A	N/A	0.233
UnionCom	0.411	0.406	0.422	0.332	0.315	0.276	0.474	0.306	0.349
Cross-m. AE	0.511	0.327	0.412	0.363	0.281	0.344	N/A	N/A	0.326
bindSC	0.713	0.584	0.475	0.387	0.254	0.262	N/A	N/A	0.412
Seurat	0.428	0.517	0.503	0.408	0.377	0.329	N/A	N/A	0.387

Values in bold show the best alignment performance, as measured by LTA.

We run SCOTv2 and SCOT using their heuristics to approximately self-tune hyperparameters. We use default parameters for other methods due to a lack of similar procedures for unsupervised self-tuning.

LTA, label transfer accuracy; SCOT, Single Cell alignment using Optimal Transport.

and SCOT on Intel Xeon e5-2670 CPU with 16 GB memory. We also run SCOTv2 on the same CPU to give comparable results with Pamona's runtimes. Figure 5 depicts that SCOT, MMD-MA, Pamona, and SCOTv2 show similar computational scaling.

5. DISCUSSION

We present SCOTv2, an improved unsupervised alignment algorithm for multiomics single-cell alignment. It extends the alignment capabilities of SCOT to data sets with cell-type representation disproportions across different sequencing measurements. It also performs alignment for single-cell data sets with more than two measurements ($M > 2$). Experiments on real-world subsampled coassay data sets and separately sampled and sequenced single-cell data sets demonstrate that SCOTv2 reliably yields high-quality alignments for a wide range of cell-type disproportions without compromising its computational scalability. Furthermore, SCOTv2's flexible marginal constraints enable it to consistently give good alignment results for both balanced and unbalanced single-cell data sets. In addition to effectively handling cell-type imbalances and multiomics alignment, SCOTv2 can self-tune its hyperparameters making it applicable in complete unsupervised settings. Therefore, SCOTv2 offers a convenient way to align multiple single-cell measurements without requiring any orthogonal correspondence information.

In this second iteration of SCOT, we have utilized the coupling matrix in a new way to find a latent embedding space. While this dimension reduction improves cell-type separation, using the coupling matrix

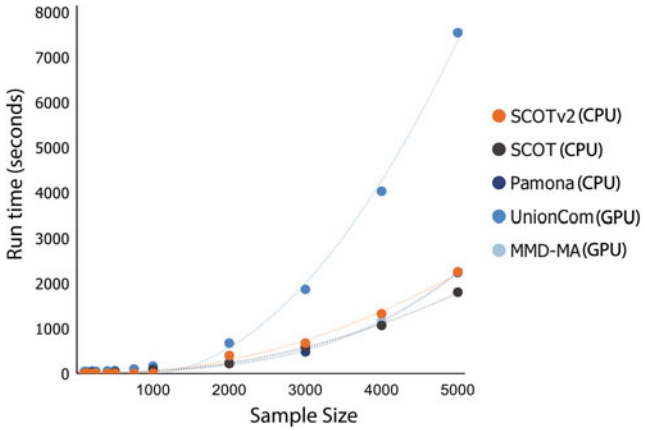

FIG. 5. Runtimes for SCOTv2, SCOT, Pamona, UnionCom, and MMD-MA as the number of samples increases.

directly may offer even more insights into interactions between the aligned domains. Future work will consider how to use the probabilities in the coupling matrix directly for downstream analysis such as improved clustering and pseudotime inference. Although SCOTv2 has runtimes that scale with other methods, it requires $O(n^2)$ memory storage for the distance matrices, which may be an issue for especially large data sets. One way to address this limitation would be to develop a procedure to align a representative subset of each domain that can be extended to the entire data set. Therefore, we will explore this direction to further improve the scalability of SCOTv2.

AUTHORS' CONTRIBUTIONS

P.D., R. Santorella, and R. Singh conceptualized the work. P.D. and R. Santorella worked on the method. P.D. set up the code and ran the analyses and M.C. helped with running the baselines. All the authors participated in the project discussions and contributed to the article preparation.

ACKNOWLEDGMENT

An earlier version of this article has been deposited to bioRxiv and published as part of the 2022 Annual International Conference on Research in Computational Molecular Biology (RECOMB).

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

R. Singh's and P.D.'s contribution to the work is supported by National Institute of Health (NIH) award 1R35HG011939-01. B.S. was partially supported by National Science Foundation (NSF) awards 1714429 and 1740741. R. Santorella is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1644760.

SUPPLEMENTARY MATERIAL

Supplementary Figure S1

REFERENCES

Alvarez-Melis D, Jaakkola TS. Gromov-Wasserstein alignment of word embedding spaces. arXiv preprint arXiv: 1809.00013, 2018; doi: 10.48550/arXiv.1809.00013

Argelaguet R, Clark SJ, Mohammed H, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature 2019;576(7787):487–491; doi: 10.1038/s41586-019-1825-8

Bonora G, Ramani V, Singh R, et al. Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and x inactivation. Genome Biol 2021;22(1):279; doi: 10.1186/s13059-021-02432-w

Cao K, Bai X, Hong Y, et al. Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics 2020;36(Suppl 1):i48–i56; doi: 10.1093/bioinformatics/btaa443

Cao K, Hong Y, Wan L. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. Bioinformatics 2021;38(1):211–219; doi: 10.13039/501100001809

Chen S, Lake BB, Zhang K. High-throughput sequencing of transcriptome and chromatin accessibility in the same cell. Nat Biotechnol 2019;37(12):1452–1457; doi: 10.1038/s41587-019-0290-0

Cheow LF, Courtois ET, Tan Y, et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. Nat Methods 2016;13(10):833–836; doi: 10.1038/nmeth.3961

Clark SJ, Argelaguet R, Kapourani C-A, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun 2018;9(1):781; doi: 10.1038/s41467-018-03149-4

Clark SJ, Argelaguet R, Kapourani C-A, et al. Integrative methods and practical challenges for single-cell multi-omics. Trends Biotechnol 2020;9(1):1–9; doi: 10.1016/j.tibtech.2020.02.013

Demetci P, Santorella R, Sandstede B, et al. Gromov-wasserstein optimal transport to align single-cell multi-omics data. BioRxiv 2020; doi: 10.1101/2020.04.28.066787

Dou J, Liang S, Mohanty V, et al. Unbiased integration of single cell multi-omics data. bioRxiv 2020; doi: 10.1101/2020.12.11.422014

González-Blas CB, Minnoye L, Papasokrati D, et al. cisTopic: cis-Regulatory topic modelling on single-cell ATAC-seq data. Nat Methods 2018;6(5):397–400; doi: 10.1038/s41592-019-0367-1

Liero M, Mielke A, Savaré G. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. Invent Math 2018;211(3):969–1117; doi: 10.1007/s00222-017-0759-8

Liu J, Huang Y, Singh R, et al. Jointly embedding multiple single-cell omics measurements. bioRxiv 2019;644310; doi: 10.1101/644310

Singh R, Demetci P, Bonora G, et al. Unsupervised Manifold Alignment for Single-Cell Multi-Omics Data. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; Virtual Event, USA. 2020; pp. 1–10.

Sjourn T, Vialard F-X, Peyr G. The unbalanced Gromov Wasserstein distance: Conic formulation and relaxation. In *Advances in Neural Information Processing Systems;* Virtual Event. 2021; pp. 8766–8779; doi: 1048550/arXiv.2009.04266

Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell 2019;77(7):1888–1902; doi: 10.1016/j.cell.2019.05.03

Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11):2579–2605.

Welch JD, Hartemink AJ, Prins JF. Matcher: Manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol 2017;18(1):138; doi: 10.1186/s13059-017-1269-0

Yang KD, Belyaeva A, Venkatachalapathy S, et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat Commun 2021;12(1):31; doi: 10.1038/s41467-020-20249-2

Address correspondence to:
Dr. Ritambhara Singh
Department of Computer Science
Brown University
115 Waterman Street
Providence, RI 02912
USA

E-mail: ritambhara@brown.edu